ChinaXiv合作期刊

基于主成分分析法搭建 A 型星有效温度的神经网络模型*

李正泽[1,2]，赵刚[1,2]

（1.中国科学院光学天文重点实验室（国家天文台），北京 100101

2.中国科学院大学天文与空间科学学院，北京 100049）

摘要：大天区面积多目标光纤光谱天文望远镜（Large Sky Area Multi-Object Fiber Spectroscopy Telescope，LAMOST，又叫郭守镜望远镜）巡天项目提供了海量恒星光谱数据，DR5 数据集中包含大量 A 型星谱线指数和有效温度的信息。机器学习算法例如可以发掘数据底层相互关系的神经网络模型广泛运用于多个学科。通过使用 DR5 数据集中的 A 型星 19 种谱线指数和有效温度数据，通过主成分分析法，给出了每种谱线指数占整个数据信息的百分比，并以此为基础，选取与有效温度关系最紧密的 12 种谱线指数数据，利用有效温度误差小于 100K 的数据训练得到有效温度的神经网络回归模型。模型在测试数据集上整体表现较好，程序给出的决定系数R^2为 0.904，平均绝对误差为 58.38K。对比相关研究的模型，测量准确度有了明显提升。此外，通过建立模型，对有效温度误差大于 100K 的原始数据重新进行测量，得到的有效温度数据绝对误差的平均值有了明显下降；同时 DR5 数据集中 A5 型恒星数据缺少有效温度参数，通过模型的测量，对这一部分数据进行了补充，提供了一定程度的参考意义。

关键字：神经网络；主成分分析；A 型星；

中图分类号：P144.2　　文献标识码：A　　文章编号：1672-7673（2020）03

根据哈佛恒星光谱分类方法，恒星的光谱可分为 O，B，A，F，G，K，M，R，S，N 等光

*基金项目：国家自然科学基金（11988101，11890694）；国家重点研发计划（2019YFA0405502）资助.
收稿日期：2019-12-12；修订日期：2020-01-11
作者简介：李正泽，男，硕士研究生.研究方向：机器学习.Email:zzli@nao.cas.cn
通讯作者：赵刚，男，研究员.研究方向：银河系结构与演化，天体丰度. Email: gzhao@nao.cas.cn

chinaXiv:202004.00031v1

谱型，对应恒星的温度依次递减， A 型星的温度区间位于 7 500 K 至 11 000 K，呈白色，有强烈的氢吸收线，并且由于温度很高，同时具有电离钙和电离镁线[1][6]。于 1993 年提出建设的 LAMOST 项目[2]，2009 年通过验收观测至今已经十余年，数据集 DR5 包括 4154 个观测天区，发布 901 万条光谱，其中包含大量的 A 型星的谱线指数数据和恒星参数数据。

相对于简单传统的回归模型，通过神经网络建立的回归模型可以更高效准确地完成任务，这要归功于神经网络模型可以捕捉非线性效应和更高阶的相互作用。对于较为复杂的数据和问题，神经网络可以挖掘出数据背后的相关性，并且给出比较令人满意的结果，在数据处理领域，以神经网络为例的众多机器学习算法已经被广泛运用于各个学科的研究之中。

包括有效温度在内的恒星参数是决定恒星光谱的重要信息，对恒星演化的研究具有重要意义[11]。对于包括有效温度在内的恒星参数的测量方法，主要有两类[8]：（1）通过将待测恒星光谱与已知参数的标准恒星光谱进行匹配，将匹配最好的模板光谱参数作为待测恒星参数。（2）类似非线性回归的方法，比如神经网络模型，利用光谱数据通过神经网络结构训练测试恒星大气参数[8]。谱线指数是包含恒星自身物理特征信息的重要参数，利用谱线指数可以进行众多的天文研究，例如：文[12]利用谱线指数数据对恒星光谱进行聚类分析研究。文[7]利用谱线指数建立人工神经网络对包括有效温度在内的恒星参数进行了测量，文中使用 LAMOST 数据训练得到的模型，预测得到有效温度的误差正态分布数学期望为-316.02，标准差为 617.36。使用 SDSS DR8 数据训练的模型结果稍好，但误差的正态分布数学期望为 88.58，标准差为 147.81。可见文中的方法还不能比较准确地给出有效温度数据，需要进一步的改进与研究。

本文使用主成分分析方法（Principal Components Analysis，PCA），运用于 LAMOST DR5 数据集中的 A 型星数据，对 19 种谱线指数数据进行相关性降维，再给出每种谱线指数占整个数据信息的百分比，以此为依据，选择与有效温度关系最紧密的几种谱线指数作为模型的输入，经过测试，选择占比最大的前 12 种谱线指数数据作为神经网络模型的输入。同时选择有效温度误差小于 100K 的数据作为输入数据，训练得到了 A 型星的谱线指数与有效温度的神经网络回归模型。通过建立的神经网络模型，给出了 8644 组有效温度误差大于 100K 的 A 型星有效温度数据，一定程度上对数据进行了改进与提升，并且通过神经网络模型对 LAMOST DR5 数据集中光谱型为 A5，缺少有效温度数据的 A5 型星数据进行了补充，给出了这些恒星的有效温度数据，提供了一定的参考意义。

1 主成分分析

如今科学研究所面临的问题日渐深入复杂，要处理的数据量也随之剧增，单纯直接处理庞大的数据已经不能满足科学研究对高效性地追求。为了从复杂繁琐的数据中提取主要信息，必须利用一些科学手段，寻找数据之间的相关性，对数据进行简化，有效减少数据的维度，但同时保证数据提供的信息极大程度地保留下来，尽量减少在这个过程中数据所携带信息的损失。主成分分析法便是为此应运而生的一种算法，现在已经成为使用最广泛的降维方法之一。

主成分分析法是一种运用十分广泛的降维方法。对于大样本多参量观测数据，它可以简捷有效地寻求参量之间的相互关系，从而对数据降维，可以去除数据噪声，消除数据沉余，使得数据更易被使用。主成分分析法的主要思想是找出数据最主要的信息、最主要的成分代替原始数据，以此达到对原始数据降维的目的，即在减少需要分析的指标的同时，尽量减少原指标所包含的信息的损失。这种方法最早被应用于社会科学的研究领域。之后随着 20 世

chinaXiv:202004.00031v1

chinaXiv:202004.00031v1

纪 60 年代计算机的兴起和发展，开始广泛运用于自然科学的研究领域[3]，于此同时，主成分分析法也开始运用于天体物理学领域，在近几年的天文研究中，文[4]利用 LAMOST 巡天光谱 DR2 数据，使用 R 语言的主成分分析工具提取各类型光谱数据的特征量，从含有大量冗余信息的光谱中提取代表恒星光谱特征的主要成分，除此之外在星系和恒星的光谱分类[6]、特征参量的挑选、活动星系核光变的研究、大样本天体红移的测量等方面，主成分分析法都有不错的表现[3]。近年来随着计算机与机器学习的飞速发展，为了克服主成分分析法的一些缺点，开发了很多主成分分析法的一些变种，比如解决非线性降维的 KPCA，解决内存限制的增量 PCA 方法(Incremental PCA)，以及解决稀疏数据降维的 PCA 方法 Sparse PCA 等。

1.1 主成分分析的数学原理

接下来在天文观测的背景下，介绍主成分分析法的数学原理，首先假设需要处理分析的数据样本由n个天体组成，每个天体对应m个观测参量，即m个特征指标，因此，观测量可以表示成矩阵$\boldsymbol{X}$，如（1）式，矩阵$\boldsymbol{X}$称之为观测矩阵，其行矢量对应同一天体的不同特征量，列矢量对应不同天体的同一特征量。

$$\boldsymbol{X} = (x_{ij})_{n*m} = \begin{vmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{vmatrix} \quad (1)$$

$$pc = eX = e_1x_{k1} + \cdots + e_ix_{ki} + \cdots + e_mx_{km} \quad (2)$$

设待求的m维特征向量为$\boldsymbol{e}$，则一个主成分pc可以表示成（2）式。同时，为了保证在降维过程中数据所携带的信息不丢失，降维后的主成分应尽可能多地体现原始观测量的信息，并且保证主成分之间互相独立。随机变量的方差可以体现随机变量所携带的信息，而不同的特征向量$\boldsymbol{e}$其方差的大小也不同，主成分分析法就是寻找使主成分pc的方差达到最大的一组特征向量$\boldsymbol{e}$。为此根据最小二乘法原理，此处的$\boldsymbol{e}$为观测矩阵$\boldsymbol{X}$的协方差矩阵$\boldsymbol{C} = (c_{jk})_{m\times m}$的正交特征矢量，其中$c_{jk}$的表达式如（3）式，$\bar{x}_j$，$\bar{x}_k$为列矢量的平均值。

$$c_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right) \quad 1 \le j,k \le m \quad (3)$$

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\ \bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik} \quad (4)$$

构造行列式方程$|\boldsymbol{C} - l\boldsymbol{I}| = 0$，其中$l$为行列式的特征根，$\boldsymbol{I}$为（$m\times m$）的单位矩阵，通过求解这个方程，可以得到特征根$l$，再求解（5）式，就能求得特征矢量$\boldsymbol{e}_i$

$$(\boldsymbol{C} - l\boldsymbol{I})e_i = 0 \quad (5)$$

方程$|\boldsymbol{C} - l\boldsymbol{I}| = 0$可以求解得到$m$个特征根，按从大到小的顺序排列，$l_1 \ge l_2 \ge l_3 \ge \cdots \ge l_m \ge 0$。每一个特征根$l_i$对应一个特征向量$\boldsymbol{e}_i$，同时对应可得第$i$个主成分$pc_i$，最大的$l_1$对应第 1 主成分。在主成分分析法中，将$l_k/\Sigma_{i=1}^{m}l_i$ 称为主成分pc_k的贡献率，将$\Sigma_{j=1}^{k}l_j/\Sigma_{i=1}^{m}l_i$称为主成分$pc_1$，$pc_2$，$pc_3$，$\cdots$，$pc_k$（$k \le m$）的累计贡献率。

1.2 主成分分析的算法流程

（1） 对样本中每个特征指标下的数据，减去该特征的平均值，即对所有样本进行中心化；

（2） 计算样本矩阵的协方差矩阵；

（3） 求协方差矩阵的特征根和特征根所对应的特征矢量；

（4） 根据特征根的大小，计算得到每个特征根对应的贡献率和累计贡献率；

（5） 用每一个特征矢量乘以样本矩阵计算得到每一个主成分，即降维后输出的新样本。

1.3 主成分分析结果

利用 LAMOST DR5 数据集给出的谱线指数、有效温度以及有效温度误差数据，给定温度为 7500K 至 11000K 提取 A 型星的数据，之后首先对数据筛选预处理，去除一些明显异常的数据，比如空值、显示为-9999 的数据，除此之外，正常情况下谱线指数都应该是正值，但是由于郭守敬望远镜流量定标没有定好，有些谱线指数的数据出现负值，因此，在这里只选取谱线指数为正值的正常数据，一共选取 53739 组 A 型星的数据。

通过主成分分析的方法对 19 种谱线指数数据（kp12, kp18,kp6, hdelta12, hdelta24, hdelta48, hdelta64, hgamma12,hgamma24, hgamma48, hgamma54, hbeta12, hbeta24, hbeta48, hbeta60,halpha12, halpha24, halpha48, halpha70）进行相关性降维，设定累计贡献率大于 90%，得到了 3 个主成分，方差分别为：15.479，1.563，1.507。因此,主成分一贡献率α=77.82%，主成分二贡献率β=7.86%，主成分三贡献率γ=7.58%。再结合主成分分析过程中得到的转换矩阵$\boldsymbol{w}$：

$$\boldsymbol{w}=\begin{vmatrix} a_1 & a_2 & \cdots & a_{19} \\ b_1 & b_2 & \cdots & b_{19} \\ c_1 & c_2 & \cdots & c_{19} \end{vmatrix} \quad (6)$$

$$a_i'=\frac{a_i}{a_1+a_2+\cdots+a_{19}} b_i'=\frac{b_i}{b_1+b_2+\cdots+b_{19}} c_i'=\frac{c_i}{c_1+c_2+\cdots+c_{19}}, \quad i=1,2,\cdots,19 \quad (7)$$

$$p_i=a_i'\cdot\alpha+b_i'\cdot\beta+c_i'\cdot\gamma \quad i=1,2,\cdots,19 \quad (8)$$

转换矩阵$\boldsymbol{w}$每一行对应新得到的一种主成分，每一列代表每种原始特征的权重大小，根据（7）式按行进行归一化，得到每种谱线指数对应 3 个主成分的权重大小a_i'，b_i'，c_i'，之后结合每种主成分的贡献率，按照（8）式计算得到每种谱线指数占整个数据信息的百分比大小p_i。见表 1。从大到小排序如下：hgamma54, hdelta64, hgamma48, hdelta48, halpha70, hbeta60, halpha48, hbeta48, kp18, hdelta24, hgamma24, kp12, halpha24, hbeta24, kp6, hdelta12, halpha12, hgamma12, hbeta12。

表 1 每种谱线指数占整个数据信息的百分比大小

Table 1 thepercentage of the entire informationfor each spectral index

kp12	kp18	kp6	hdelta12	hdelta24	hdelta48	hdelta64	hgamma12	hgamma24	hgamma48
4.06%	4.96%	2.31%	1.96%	4.71%	7.50%	8.29%	1.91%	4.38%	8.22%
hgamma54	hbeta12	hbeta24	hbeta48	hbeta60	halpha12	halpha24	halpha48	halpha70	
8.92%	1.55%	3.23%	5.70%	6.29%	1.94%	3.70%	6.21%	7.47%	

2 搭建神经网络模型

本文使用的机器学习模型是多层感知器（Multilayer Perceptron， MLP），即神经网络模型[5][10]，在 Python 环境下提供了多种机器学习算法，其中 sklearn.neural_network 模块

chinaXiv:202004.00031v1

提供多层感知器回归算法，即 MLPRegressor[9]。多层感知器顾名思义，由多个层构成，包括一个输入层和可以规定数量的多个隐藏层以及一个输出层，隐藏层的加入增强了模型的表达能力，但同时也使得模型变得更加复杂，对于输出层的神经元来说，可以有不止一个输出。

神经网络模型设置了两个隐藏层，每个隐藏层包含 100 个节点，多层感知器回归算法 MLPRegressor 中可选择的激励函数有 4 种，分别是 identity，logistic，tanh，relu，分别测试了这 4 种激励函数下模型的表现，如表 2。由表 2 可以看出，选择 identity 和 relu 时模型表现比较好。选择 relu 时模型表现更好，并且选择 relu 时模型训练速度较快，效率较高。因此，搭建神经网络模型的激励函数设置为 relu。但是选择 relu 作为激励函数时有一个缺点，可能会造成神经元坏死，为了避免这种情况发生，在这里网络的学习速率设置的较小，避免权重突然更新过多，导致神经元彻底关闭。

表 2 不同激励函数下多层感知器的表现

Table 2 the performance of MLP byusing different Activation function

	identity	logistic	tanh	relu
Score	0.886	<0	<0	0.904
Mean absolute error	70.02K	large	large	58.38K
Standard deviation	59.22K	large	large	60.81K

经过测试，梯度下降函数选择在相对较大数据集上效果较好的 adam，此时模型运算效率较高并且结果较好。设置正则化系数 alpha 则是为了避免过拟合的发生，设置为 0.001，同时保证模型的运行结果较好。最大训练迭代次数 max_iter 经过测试设置为 4000。除此之外其他参数设定为默认设置。

2.1 选择输入参数

图 1 是郭守敬望远镜提供的有效温度的绝对误差分布图，选取有效温度误差小于 100K，共计 45095 组数据建立模型，其中随机选取 80%的数据作为训练数据，20%的数据作为训练之后的测试数据。通过主成分分析法给出了 19 种谱线指数占整个数据信息的百分比大小排序，据此，选择与有效温度关系最紧密的几种谱线指数作为神经网络模型的输入，按照信息占比从大到小的顺序依次选择 1 种到全部 19 种谱线指数作为神经网络模型输入。测试不同指标数量下模型的表现，建立模型之后 score 命令可以给出模型的评分，即模型对全部数据的预测结果的决定系数R^2，具体计算公式见（9）式，其中，U为残差平方和；y_t为真实的数据；y_p为预测的数据；V为总平方和；$\bar{y}_t$为真实数据的平均值。决定系数R^2越接近 1 表示模型与数据匹配越好[9]。

$$R^2 = 1 - \frac{U}{V}\text{其中}U = \sum_{i=1}^{n}(y_t - y_p)^2\ V = \sum_{i=1}^{n}(y_t - \bar{y}_t)^2 \quad (9)$$

表 3 与图 2 是以模型的评分为标准给出的结果。可以看出，选取包含信息最多的前 12 种谱线指数数据时，模型的评分最高，模型表现最好，因此，选取前 12 种谱线指数，即 hgamma54，hdelta64，hgamma48，hdelta48，halpha70，hbeta60，halpha48，hbeta48，kp18，hdelta24，hgamma24，kp12 作为神经网络模型的输入。

chinaXiv:202004.00031v1

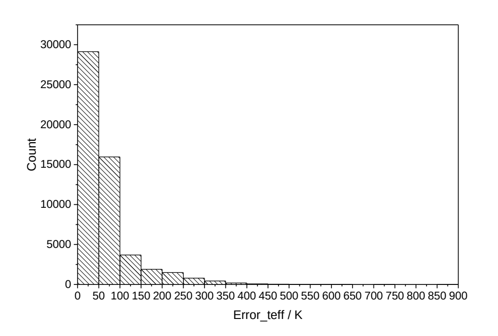


图 1　有效温度绝对误差分布图

Fig 1 Absolute error distribution diagram effective temperature

表 3 不同指标数量下模型的评分

Table 3the model score for different number of features

1	2	3	4	5	6	7	8	9	10
0.498	0.597	0.703	0.743	0.732	0.743	0.759	0.761	0.874	0.882
11	12	13	14	15	16	17	18	19	
0.880	0.904	0.873	0.880	0.895	0.887	0.897	0.888	0.872	

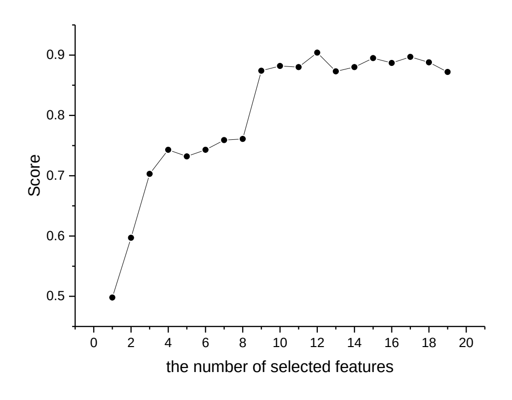


图 2 模型评分随指标数量的变化

Fig.2 the relationship between score and the number of selected features

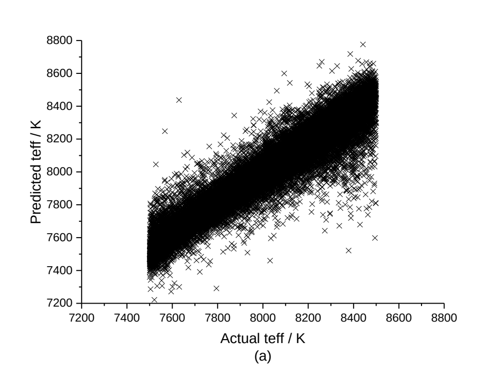
2.2　建立模型

2.2.1　模型在训练数据集上的表现

在 80%的训练数据集上，用得到的神经网络模型对有效温度进行了预测，如图 3（a），训练数据集 36076 个数据点整体分布在相对集中的区域，个别数据偏离较大，除此之外，由图 3（b）可以看出，随着有效温度变大，误差存在一个轻微的下降趋势，文[7]对于这个现象的解释是可能因为人工神经网络内部的机制的原因,考虑到郭守敬望远镜数据本身对于早型星的恒星参数测量并不准确，所以有可能是数据本身的影响造成的，有待进行更加深入的讨论。经过计算绝对误差的平均值为 58.12K，标准差为 60.99K，结合测试数据集上的预测结果，两者的平均绝对误差和标准差的结果基本一致，由此可以表明，神经网络模型并没有发生过拟合。

chinaXiv:202004.00031v1

chinaXiv:202004.00031v1



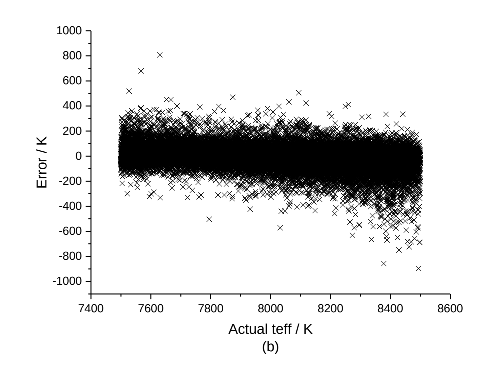
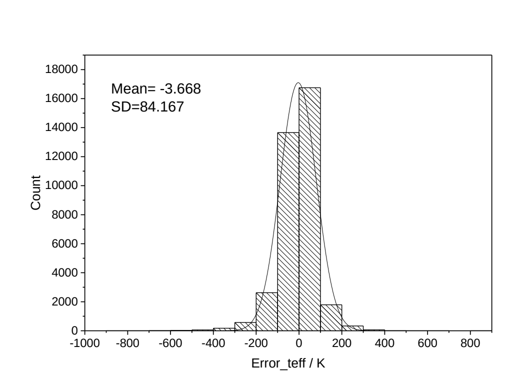


图 3 神经网络回归模型在训练数据集上的预测结果

Fig.3 the results of forecast by neural networkontraindataset



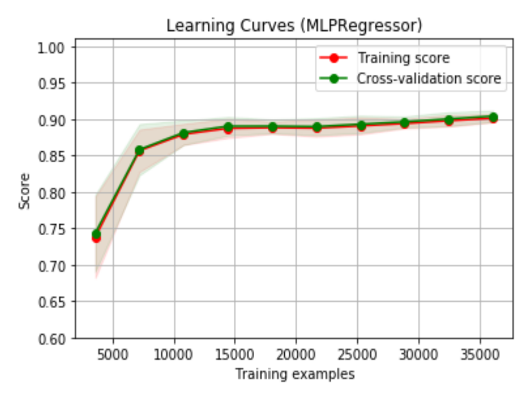


图 4 训练数据集上有效温度的误差分布图图 5 神经网络学习曲线

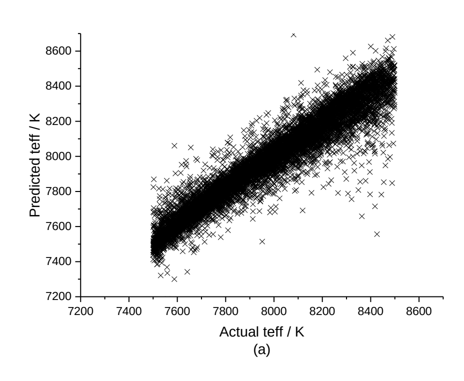
Fig.4 error distribution diagram of effective temperature on train data set

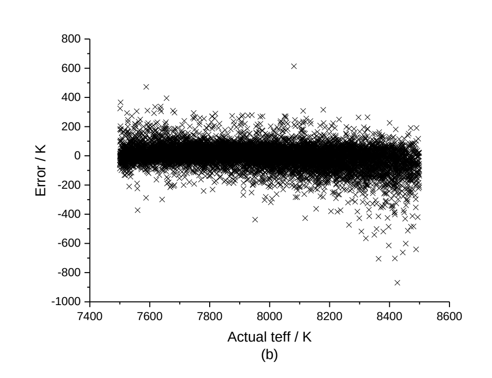
Fig.5 learning curves

图 4 给出了误差分布图及其拟合的正态分布曲线，正态分布的数学期望为-3.668，标准差为 84.167。图 5 是神经网络模型的学习曲线，从图中可以看出，随着训练样本数量的增加，训练得分（图中红线部分）快速增加，达到饱和之后趋于水平。测试得分（图中绿线部分）与训练得分变化趋势一致，但是并没有出现训练得分较高，测试得分较低或者测试得分达到某一值后迅速下降，即过拟合的情况。除此之外，训练得分与测试得分都处于较高的水平，因此神经网络模型并没有欠拟合。整体来看，模型的学习曲线收敛且误差较小，是一条比较理想的学习曲线。

2.2.2 模型在测试数据集上的表现

对于神经网络回归模型，程序给出的评分为 0.904，图 6 是在测试数据集上得到的有效温度的预测结果，其中，由图 6（a）可以看出，预测有效温度值与实际有效温度值成正比，整体预测结果较好，绝对误差的平均值为 58.38K，不足 A 型星有效温度的百分之一，标准差为 60.81K，但是还是存在个别预测数据与实际数据偏离较大；图 6（b）给出了模型的误差变化趋势，可以看出，误差围绕在纵坐标轴 y=0 上下，个别数据还是出现了较大的偏离，除此之外，还能够看出误差有一个轻微的下降趋势。图 6（c）给出了误差的分布图及其拟合的正态分布曲线，可以看出与训练数据集上的结果一致，误差主要集中在 100K 以内，正态分布拟合的数学期望为-3.366，标准差为 84.229。可见模型的有效温度预测准确度相比文[7]建立的模型有了很大的改进与提升。





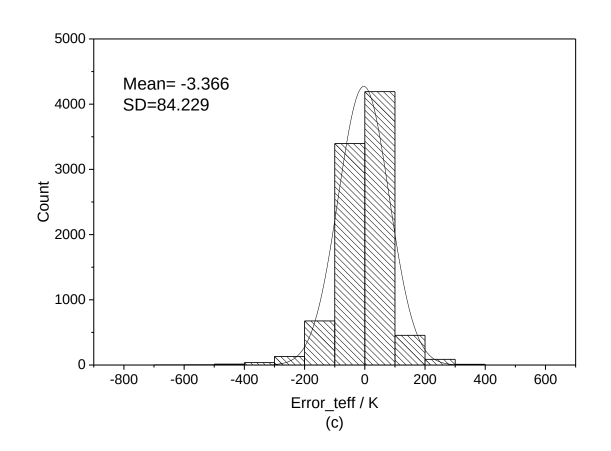


图 6 神经网络回归模型在测试数据集上的预测结果

Fig.6 the results of forecast by neural networkontestdataset

chinaXiv:202004.00031v1

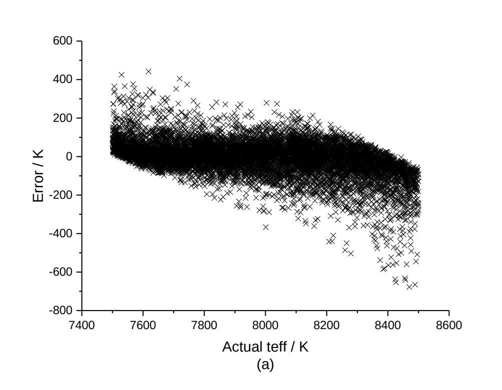
2.3 不同模型比较

支持向量机和神经网络模型都可以解决非线性的回归问题，通过 sklearn.svm 中的 SVR 模块，建立了一个支持向量机回归模型（Support Vector Regression, SVR）与前文的神经网络模型进行了对比，见表 4。此外还建立了一个决策树回归模型（Decision Tree Regression,DTR），选取 80%的数据作为训练数据，20%的数据作为测试数据，为了防止严重过拟合的发生，经过测试决策树回归模型的最大深度设置为 6。查看决策树回归模型在两个数据集上的结果，此时在训练数据集上绝对误差的平均值为 65.10K，标准差为 61.74K，在测试数据集上绝对误差的平均值为 66.76K，标准差为 62.83K，因此，模型没有发生过拟合。表 4 给出了 3 种模型在测试数据集上的结果对比。可以看出，神经网络模型在评分和误差方面比支持向量机、决策树回归模型较好的结果。图 7（a）和图 7（b）分别给出了支持向量机和决策树回归模型在测试数据集上的误差变化，前文提到神经网络模型随着有效温度的变大，误差存在一个轻微的下降趋势，从图 7（a）支持向量机模型整体来看，误差也存在一个下降的趋势，尤其是 8200K 到 8500K 之间，误差有明显的下降趋势，因此，产生这个现象的原因可能不单单是神经网络内部的原因，也可能与数据本身有关。

表 4 模型的比较

Table 4 thecomparisonofdifferentmodels

	DTR	SVR	MLP
Score	0.890	0.882	0.904
Mean absolute error	66.76K	67.23K	58.38K
Standard deviation	62.83K	68.64K	60.81K



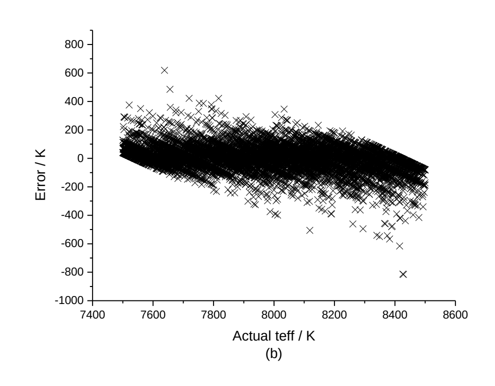


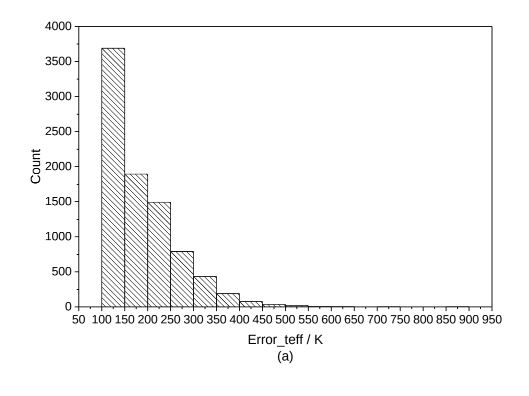
图 7 支持向量机和决策树回归模型在测试数据集上的预测结果

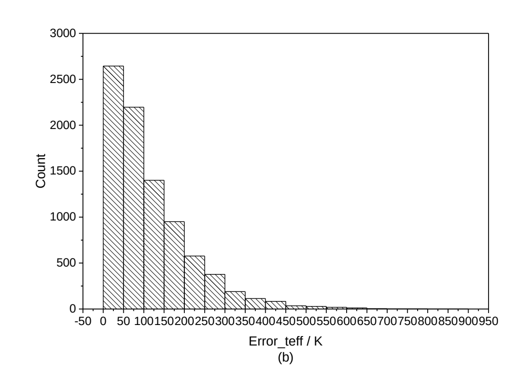
Fig.7 the results of forecast by SVR and DTR

3 神经网络模型的应用

3.1 对有效温度误差较大的数据进行改进

选取了 LAMOST DR5 数据集中包含有有效温度、有效温度绝对误差以及 19 种谱线指数数据的 A 型星数据，共计 53739 组，使用其中有效温度误差小于 100K 共 45095 组数据建立了神经网络模型。通过建立的神经网络模型对有效温度误差大于 100K 的 8644 组数据，使用其谱线指数数据进行了计算预测，给出了有效温度值，对数据进行了改进与提升，提供了一定程度的参考价值。对于 LAMOST DR5 数据集中有效温度绝对误差大于 100K 的数据，图 8（a）是有效温度绝对误差的分布图，图 8（b）是通过模型的预测得到的有效温度的绝对误差分布图。对于 LAMOST 给出的有效温度绝对误差，平均值为 185.10K，标准差为 78.79K；神经网络模型给出的有效温度绝对误差平均值为 115.24K，标准差为 104.88K。可以看出，有效温度绝对误差平均值有明显下降，对于有效温度数据一定程度上有所改进与提升。





chinaXiv:202004.00031v1

图 8 有效温度绝对误差与模型预测得到的有效温度绝对误差分布图

Fig.8absolute error distribution diagram of effective temperature for LAMOST and prediction

3.2 对 LAMOST 缺少的 A5 光谱型恒星有效温度数据进行补充

DR5 数据集一共给出了 40 多万条 A 型星光谱，但明确给出有效温度数据的 A 型星只有 8 万多颗，这其中还包括很多误差非常大的数据。对于有效温度的测量，通过神经网络模型可以使用谱线指数的数据，更加自动高效地进行测量，一定程度上弥补这部分数据的缺失，提供一定的参考意义。依据哈佛天文台的恒星光谱分类系统，除了分为 O、B、A、F、G、K、M、R、S、N 几个光谱型之外，对于每种光谱型还可以分为 10 个次型，用数字 0 到 9 表示，并且对应恒星的温度依次下降[6]。考虑到模型使用有效温度 7500K 到 8500K 的数据训练建立的，这里选取温度区间相近的光谱型恒星，以 A5 型恒星数据为例[6]，LAMOST 提供了谱线指数数据且分类为 A5 型的恒星一共有 470 组，基本没有给出有效温度的数据。考虑到流量定标没有定好，导致谱线指数出现负值的情况，选取其中每种谱线指数都大于 0 的数据，通过神经网络模型给出了这些恒星的有效温度数据，表 5 展示了其中一小部分结果，包括观测号（obsid），赤纬（Dec），赤经（Ra）和预测得到的有效温度（teff）。根据 MK 分类系统的光谱型与有效温度之间的关系[6]，对于 A5 型恒星来说，光度级为Ⅰ（超巨星），即 A5Ⅰ型恒星的有效温度为 8610K；光度级为Ⅴ（主序星），即 A5Ⅴ型恒星的有效温度为 8180K，光度级Ⅵ（亚矮星）的恒星的有效温度更低。考虑到观测数据的分类以及谱线指数数据都可能不准确，预测得到的 A5 型恒星的有效温度基本符合上述范围。

表 5 预测得到 LAMOST DR5 数据集中 A5 型恒星有效温度

Table 5predicted effective temperature of A5 type star in LAMOST DR5 data set

obsid	Ra	Dec	teff	obsid	Ra	Dec	teff
4113006	51.65291	52.66395	8571.45	555314004	91.01701	20.9542681	7987.72
4607082	36.6579	56.944608	7839.64	557706227	235.2796	0.8151459	6991.45
6808044	98.74538	28.21549	7154.94	565210158	273.0517	1.957477	8497.29
15004223	57.59596	50.719356	8172.05	583707042	281.6488	0.500164	7415.48
15010203	54.65725	49.709679	8807.14	583707045	281.6511	0.224461	7120.40
38501228	60.50061	47.978225	7450.72	573311118	275.7202	-0.859244	8354.52
38504105	62.32311	50.106506	7339.61	573504037	281.2724	7.070487	8316.87
…	…	…	…	…	…	…	…
506112038	309.9581	43.619166	8170.19	289916227	83.09115	37.684256	8311.47
506113076	311.0851	41.94965	8640.66	250705140	306.9164	37.379419	8564.75

4. 结论

通过 LAMOST DR5 数据集提供的 A 型星 19 种谱线指数与有效温度数据，通过主成分分析法进行相关性降维，根据每种谱线指数占整个数据信息的百分比，经过测试选择了与有效温度关系最紧密的 12 种谱线指数作为输入数据。筛选有效温度误差小于 100K 的数据建立了神经网络回归模型，模型在测试数据集上表现良好，评分为 0.904，平均绝对误差为 58.38K，标准差为 60.81K。对比相关研究的模型，准确度有了很大的提升。通过有效温度神经网络回归模型对 LAMOST 提供的有效温度误差大于 100K 的数据进行了预测，经过模型预测得到的有效温度数据的绝对误差平均值有明显的下降，一定程度上对这部分数据进行了改进与提升，

chinaXiv:202004.00031v1

ChinaXiv合作期刊

此外,LAMOST DR5 数据集提供了大量的 A 型星数据，但绝大部分缺少有效温度数据，通过神经网络模型可以实现高效自动较为准确地给出这部分数据,以光谱型为 A5 的恒星数据为例,对 LAMOST 缺少有效温度的 A 型星数据进行了弥补与补充，提供了一定的参考意义。

包括 A 型星在内的早型星的恒星参数不容易测量得到，LAMOST 巡天项目提供了海量的光谱观测数据,其中包括大量的 A 型星数据，但包括有效温度在内的恒星参数数据却非常缺乏。通过本文方法验证了建立神经网络模型利用谱线指数预测有效温度的方法是有效可行的,同时该方法能够自动高效地测量有效温度,并且测量的准确度相比于前人建立的模型有了很大的改进与提升。

致谢：本文工作由国家自然科学基金（11988101，11890694）和国家重点研发计划（2019YFA0405502）资助。

参考文献

[1]Ledrew Glenn. The Real Starry Sky [J]. Journal of the Royal Astronomical Society of Canada, 2001, 95: 32.

[2] Gang Zhao, Yongheng Zhao, Yaoquan Chu, Yipeng Jing, Licai Deng. LAMOST Spectral Survey [J]. RAA, 2012, 12, 723.

[3] 李成，孔旭，程福臻. 主成分分析法在天体物理中的应用[J]. 天文学进展,2001(01):9-16.

[4] 陈淑鑫,罗阿理,孙伟民.R 语言应用于 LAMOST 光谱分析初探[J].天文研究与技术,2017,14(03):363-368.

[5] 李丽丽,张彦霞,赵永恒,杨大卫.人工神经网络在天文学中的应用[J].天文学进展,2006(04):285-295.

[6] 覃冬梅,胡占义,赵永恒.一种基于主分量分析的恒星光谱快速分类法[J].光谱学与光谱分析,2003(01):182-186.

[7] 谭鑫,潘景昌,王杰,罗阿理,屠良平.基于神经网络的线指数恒星大气物理参数测量方法[J].光谱学与光谱分析,2013,33(06):1701-1705.

[8] 潘亚春,屠良平.基于神经网络的恒星大气参数自动测量[J].辽宁科技大学学报,2009,32(01):21-26.

[9] Pedregosa et al. Scikit-learn: Machine Learning in Python [J]. JMLR, 2011,12(Oct):2825−2830.

[10] Bailer-Jones C A L, Gupta R, Singh H P. An introduction to artificial neural networks [J]. Automated Data Analysis in Astronomy, 2002, p.51.

[11]李宗伟,肖华.天体物理学[M].北京:高等教育出版社,2000:1

[12] 王光沛,潘景昌,衣振萍,韦鹏,姜斌.基于线指数特征的海量恒星光谱聚类分析研究[J].光谱学与光谱分析,2016,36(08):2646-2650.

Neural network model of the effective temperature for A type star based on principal component analysis

Li Zhengze[1,2], Zhao Gang[1,2]

chinaXiv:202004.00031v1

ChinaXiv合作期刊

(1.Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101
2. School of Astronomy and Space Science, University of Chinese Academy of Sciences, Beijing 100049)

Abstract: The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) has provided bulk of stellar spectra data. DR5 catalogue contains plenty ofspectral indices and effective temperature of A type stars. Recently machine learning algorithms such as Neural network model which can be used to explore the deep relationship between different data have been widely used in various disciplines. In this paper with19 spectral line indices and effective temperature of A type star from LAMOST DR5 data set. Through Principal component analysis (PCA), we present thepercentage of the entire informationfor each spectral index and 12 spectral line indices which are most closely related toeffective temperature are selected as an input to establish a Neural network model for effective temperature, meanwhile the absolute error of effective temperature for these input data are less than 100K. The model performs well overall on the test data set. The coefficient of determinationR^2given by the program is 0.904 and an average absolute error of 58.38K.Compared with related research model, the measurement accuracy has been significantly improved.Furthermore, fortheraw data which have absolute error more than 100K, we remeasureeffective temperature via our model and the average absolute error of the new effective temperature data has decreased significantly. Besides LAMOST DR5 catalogue barely have effective temperature of A5 type star, we make up these missing data. Thiswork provides a certain degree of reference significance.

Key words: Neural network; A type star; principal component analysis

chinaXiv:202004.00031v1